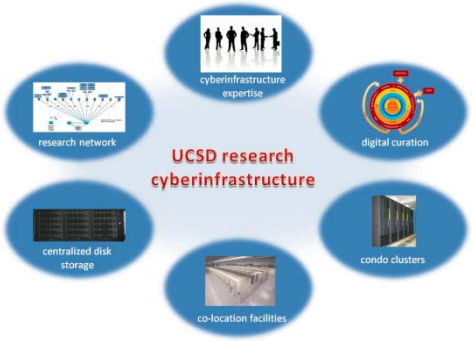
Apr 3rd, 12:00 AM - 10:35 AM

Morning Address, Part 1: "UCSD's Research CyberInfrastructure (RCI) Program: Enabling Research Thru Shared Services"

Richard Moore
University of California - San Diego

UCSD's Research CyberInfrastructure (RCI) Program:
Enabling Research Thru Shared Services



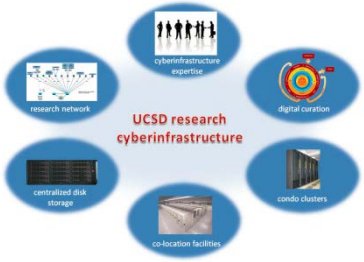
eScience Symposium

April 3, 2013

Richard Moore

University of California San Diego

San Diego Supercomputer Center



A brief history of UCSD's RCI Program …

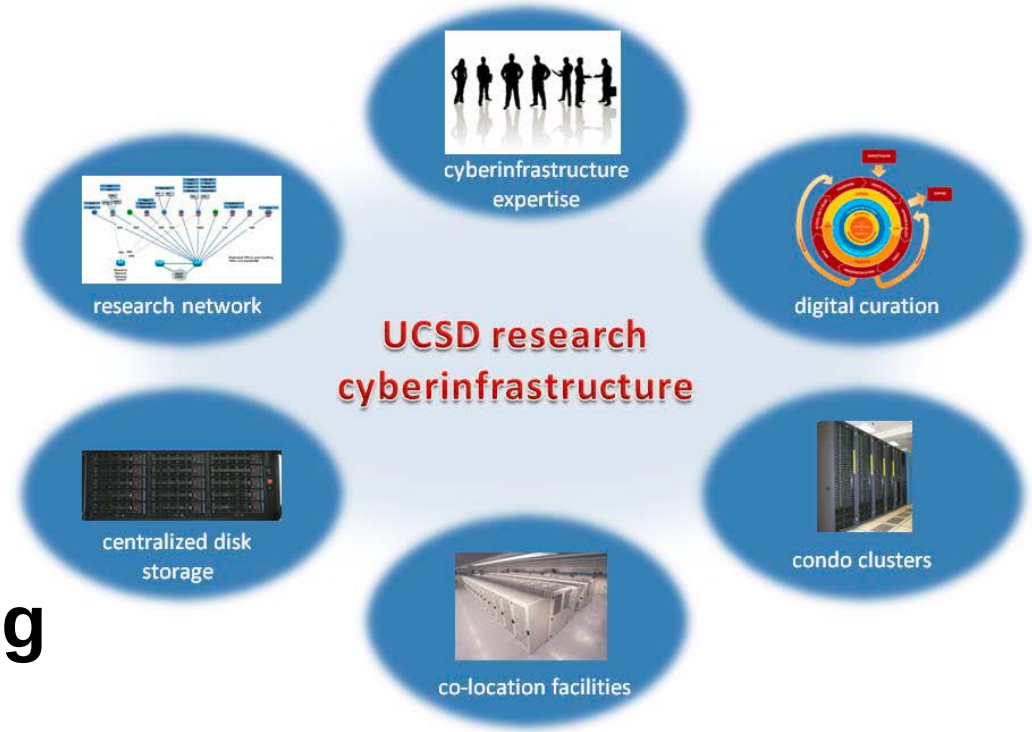
- **2008- April 2009: *Research Cyberinfrastructure Design Team (RCIDT)***
  - Broad campus participation
  - Chaired by Mike Norman and Phil Papadopoulos
  - Campus-wide survey of research cyberinfrastructure needs (2008)
  - RCIDT issued *Blueprint for the Digital University* (*http://rci.ucsd.edu/_files/RCIDTReportFinal2009.pdf*)
- **2009- April 2010: *CyberInfrastructure Planning & Operations Committee*** *(CIPOC)* developed a business plan with recommendations
  - A principle of shared costs between PIs and campus RCI investments
- **2011-Present: *RCI Oversight Committee* charged to implement RCI**
  - January 2011 - CIPOC business plan accepted, oversight committee charged
  - Broad campus representation
  - Chaired by Mike Norman and Mike Gilson

Why UCSD Is Investing in RCI

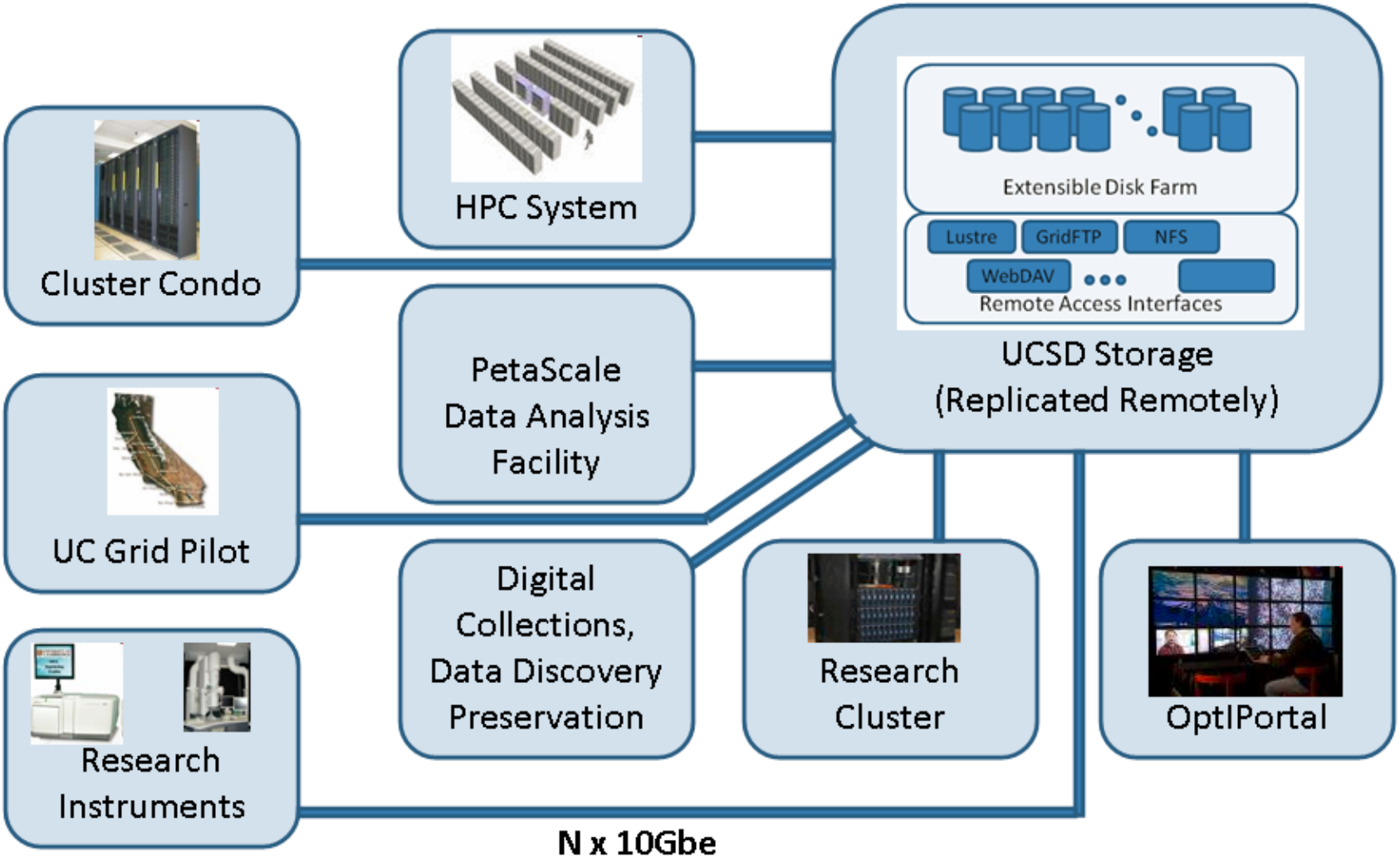
- **Increase competitiveness of UCSD researchers**
- **Realize cost efficiencies and improve service via economies of scale and shared services**
- **Preserve UCSD's digital intellectual property**
- **Save energy/$ and effectively use data center capital investments (colocation)**

Elements of UCSD's Integrated Research CyberInfrastructure

- **Data Center Colocation**
- **Networking**
- **Centralized Storage**
- **Data Curation**
- **Research Computing**
- **Technical Expertise**



A Data-Centric View of UCSD's RCI: From the 2009 Blueprint report



RCI Program is "by campus, for campus"

- **RCI priorities driven by researcher requirements**
- **Oversight Committee represents all campus units and sets strategic directions and oversees implementation**
- **Implementation partners from across campus**
  - Administrative Computing & Telecommunications
  - Calit2
  - San Diego Supercomputer Center
  - UCSD Libraries

RCI is rolling out production services for UCSD researchers

Service	Status	Lead/contact for service
Colocation	Production	Matt Campbell (SDSC) mattc@sdsc.edu
Networking	Production	Valerie Polichar (ACT) vpolichar@ucsd.edu
Research Computing	Production March 2013	Jim Hayes (SDSC) jhayes@sdsc.edu
Centralized Storage	Initial Production Spring 2013; Expanded Services thru 2013	Wilfred Li (SDSC) wilfred@sdsc.edu
Data Curation	Completing pilots; production FY13-14	David Minor (Libraries) dminor@ucsd.edu
Technical Expertise	Not planned as formal RCI service; expertise distributed across departments and projects	

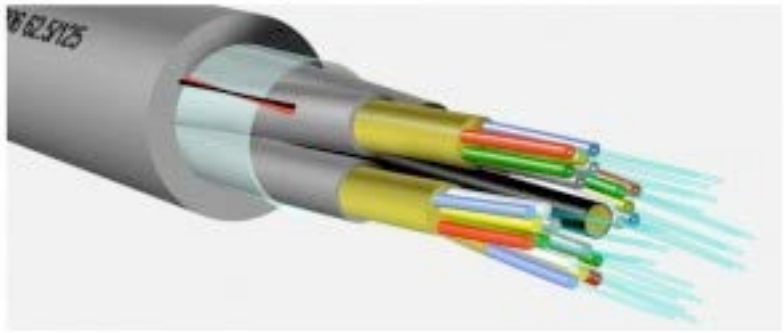
Colocation – in production

- **Host IT equipment in energy-efficient, manned data center**
  - SDSC's 18kft2, 13MW datacenter
  - Standard rack space, secure facility, seismic protection
  - 24/7 operations staff provide facility oversight and emergency "remote hands" hardware assistance

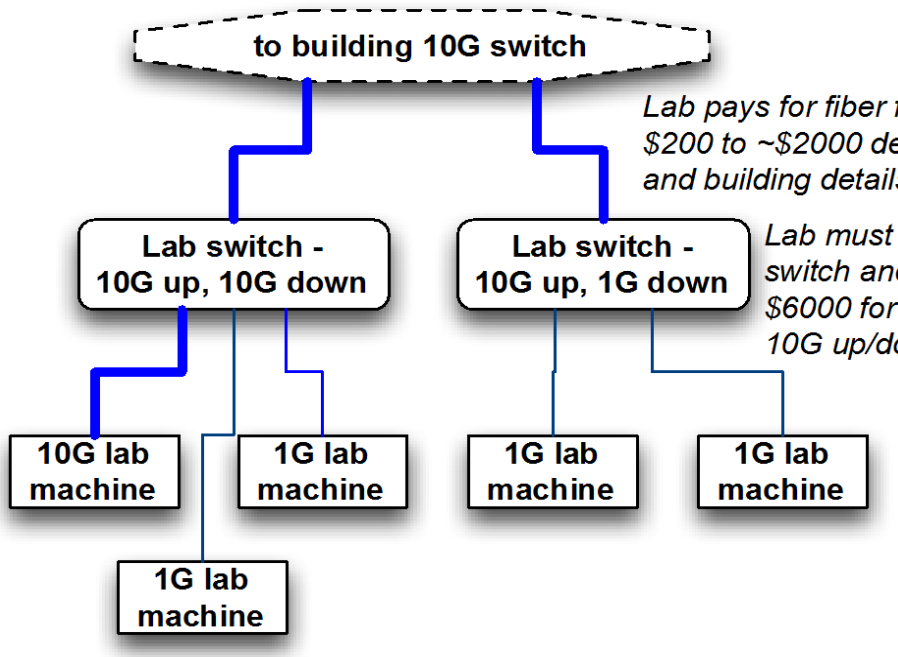
- **RCI supplements rack rate: user pays $2500/rack/year**
- **NGN *may* cover basic networking costs; evaluated case-by-case**
- **Up to 10 Gb/s networking fabric connectivity available, both thru SDSC aggregation fabric and into CENIC**
- **UPS and generator capabilities available**
- **Cage and locked rack options available for security/compliance**

UCSD's High Performance Network

- **40G campus Internet connectivity**
- **10G layer 2 connection grows to 100G this year**
- **Redundant 10G campus backbone**
- **10G+ to most campus & SIO research buildings**
- **10G or greater to research labs on request**



High-Performance Networking in the Research Lab



- **NGN3 supports 10G to the building switch & building switch 10G optics pointing towards the lab**
- **Department/lab pays for "last 100 feet" connectivity to use capability – fiber & lab switch**

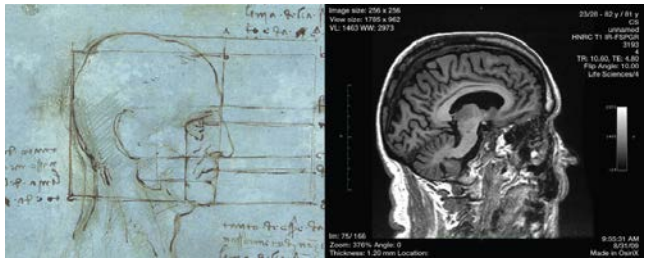
Sites requiring dedicated 1G or 10G pipes across campus (or to Internet2 or NLR) will incur specific additional costs, hard to estimate generally

Research Computing (in production)

- **RCI is evolving SDSC's Triton system to the "*Triton Shared Computing Cluster*" (TSCC)**
- **Condo model: Researchers purchase compute nodes which are operated as part of shared cluster for 3-4 years**
  - PI buys hardware & modest ops fee
  - Lower operations cost than local PI cluster; larger-scale resource available (core count and capacity); professionally-managed
- **Hotel: Purchase time by the core-hour; shared queue**

Data Curation – in pilot (production FY13-14)

- **Completing a two-year pilot phase**
  - How do lab personnel work with librarians to curate their data?
  - How much work is required to curate data and what are options?
  - What is a sustainable business model for curation within RCI project?
- **Five representative programs across UCSD selected as pilots**
  - The Brain Observatory (Annese)
  - Open Topography (Baru)
  - Levantine Archaeology Laboratory (Levy)
  - SIO Geological Collections (Norris)
  - Laboratory for Computational Astrophysics (Wagner)



- **Using existing tools whenever possible**
  - Storage at SDSC, campus high-speed networking, Digital Asset Management System (DAMS) at UCSD Libraries, Chronopolis digital preservation network
- **Also, develop Data Management Plan tools and provide training**
- **Anticipate production curation services in FY13-14**

Data Management Plans

- **Resources and contacts available to UCSD researchers**
- **Examples from submitted proposals**
- **Guidance, tips and recommendations for DMP preparation**
- **UCSD-centered version of DMP Tool**

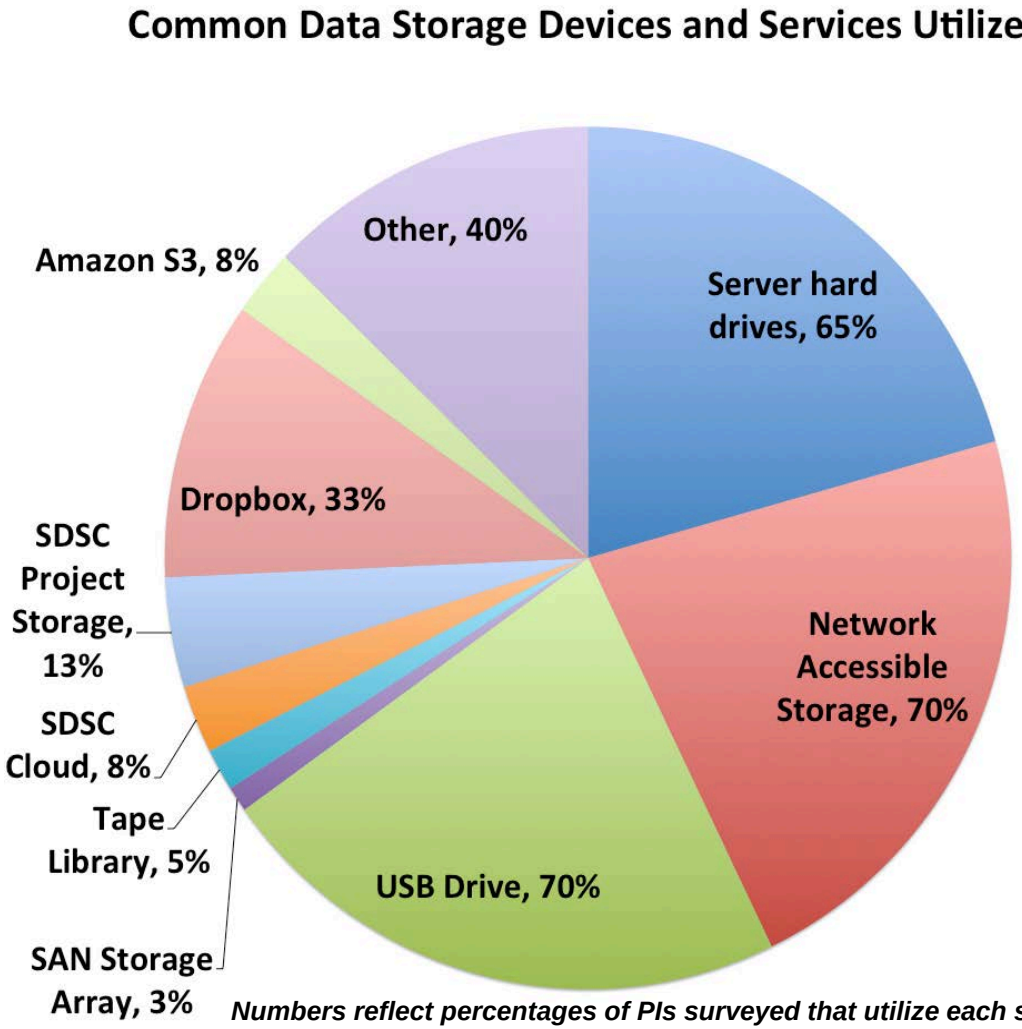
http://rci.ucsd.edu/dmp/index.html

Centralized Storage (phased production thru 2013)

- **Completing interviews of a broad sample of ~50 representative PIs to understand technical and cost requirements (report soon)**
- **Identify common needs, and define sustainable RCI business model with strong adoption**

- **Anticipate production centralized storage services in CY13**
  - **RCI Network-Attached Storage (RCI/NAS) available Spring CY13**
  - **Further services to be rolled out throughout the year, based on requirements analysis**

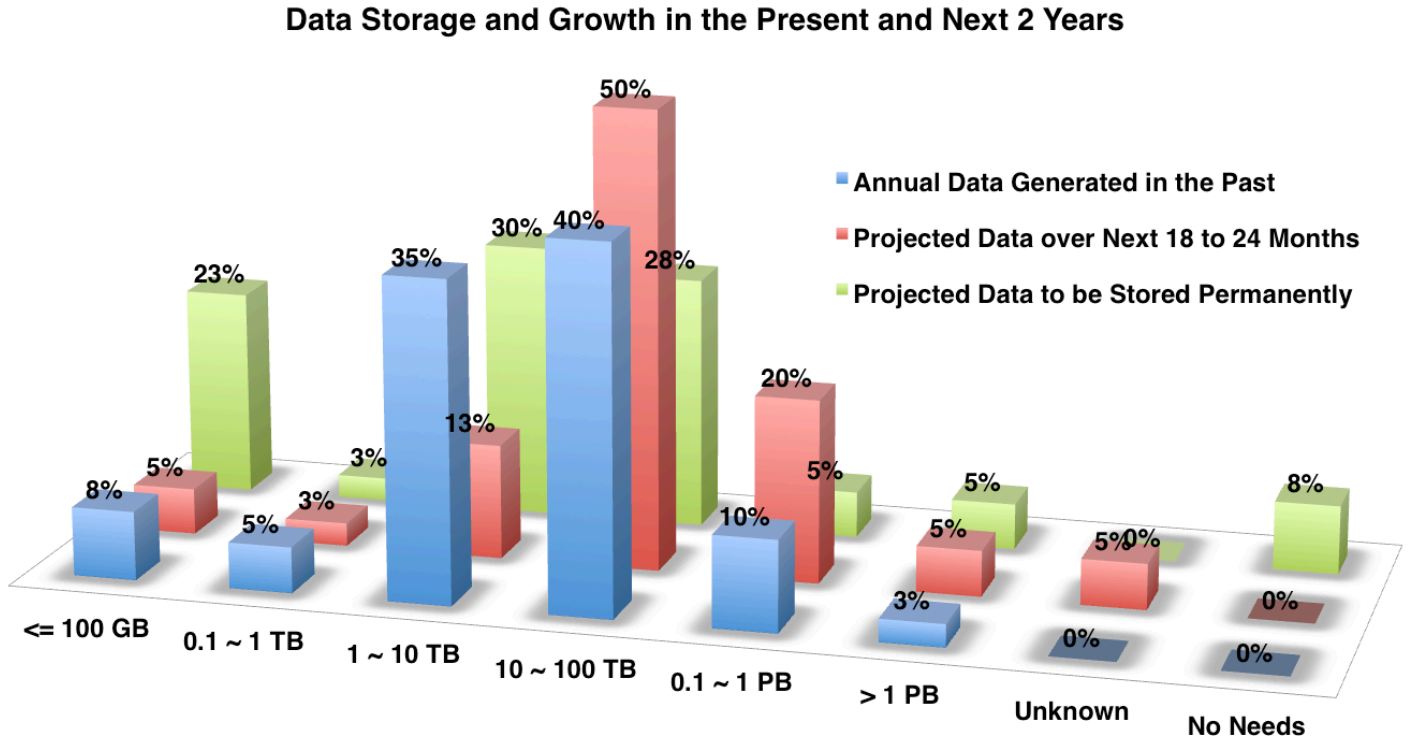
PI Interview Responses: How do You Handle Data Storage/Backup?



Numbers reflect percentages of PIs surveyed that utilize each solution ; Individual PIs use multiple solutions, so %'s add up to >100%.

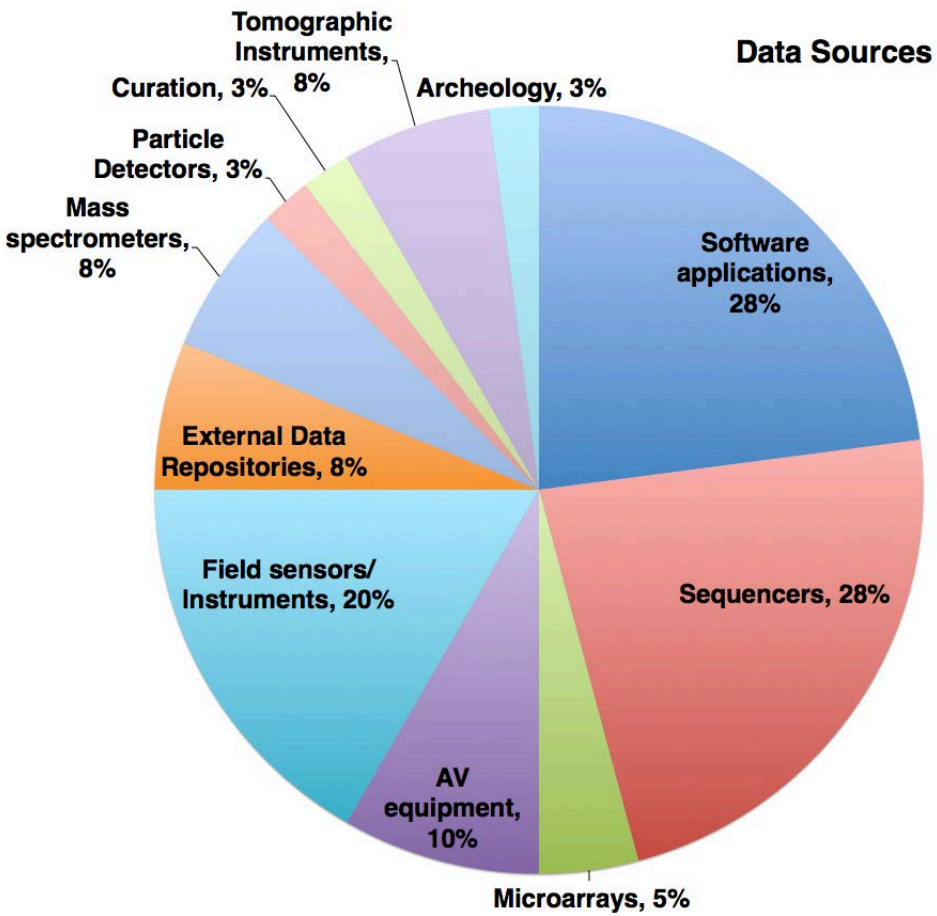
- **Storage Devices**
  - Network accessible storage (NAS), USB and server local drives dominate
  - Use of Dropbox for sharing
  - Others use Google Drive, Hadoop, XSEDE, SDSC co-location
- **Backup modes**
  - Replicated copies in two NAS
  - A copy in the NAS,
  - A copy in local hard drive (laptop/workstation),
  - And a copy in a USB drive
  - Maybe a copy in email/Dropbox
- **Problems:**
  - Out of sync
  - Lost track of its location
  - Lost version control
  - High cost of recovery

PI Interview Responses: How Much Storage Do You Need: Now, Future, Permanently?



- For PIs interviewed, current needs 1-1000TB
- Increasing in future
- Perceptions of permanent storage interesting – none for some, intermediate for many, large for a few

PI Interview Responses: Where is Your Data Coming From?

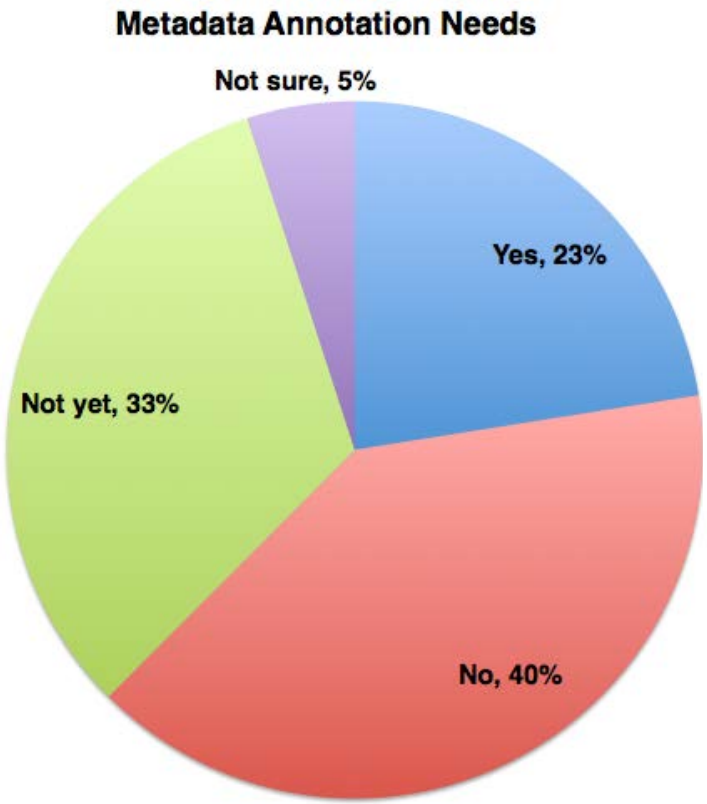


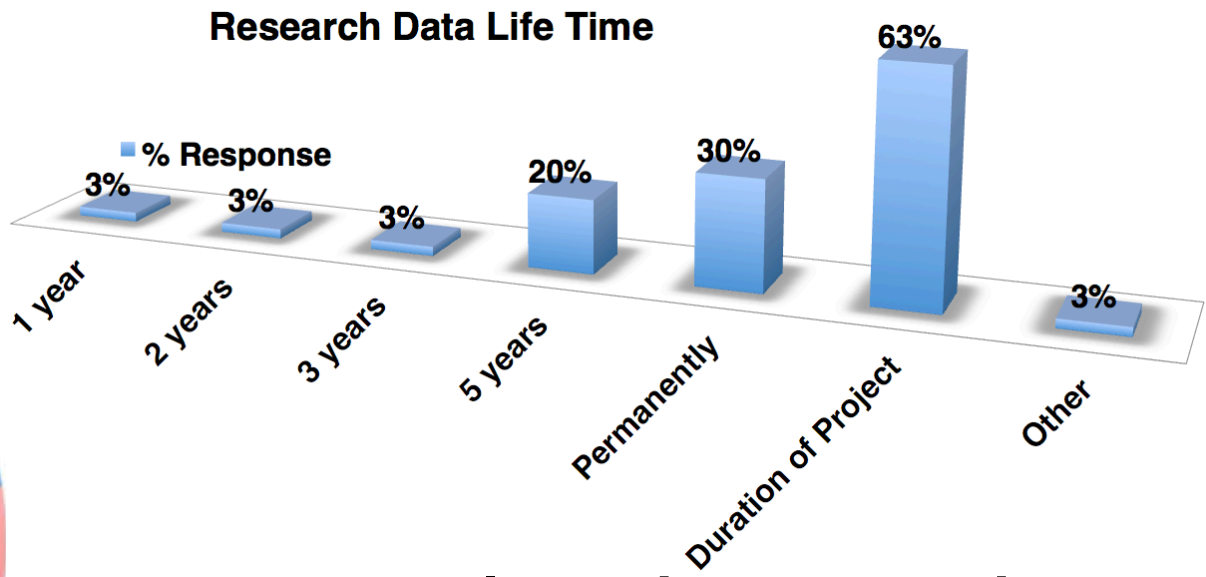
Numbers reflect percentages of PIs surveyed that utilize each solution ; Individual PIs use multiple solutions, so %'s add up to >100%.

- **Indicates use cases for connectivity requirements**
- **Data sources:**
  - ~50% campus instruments
  - ~30% simulations (XSEDE, campus, lab systems)
  - ~20% field instruments
  - ~15% other external sources
- **%'s reflect PIs, not data volume**

Interview responses: Metadata and Retention Requirements

Do you need metadata annotation capability?





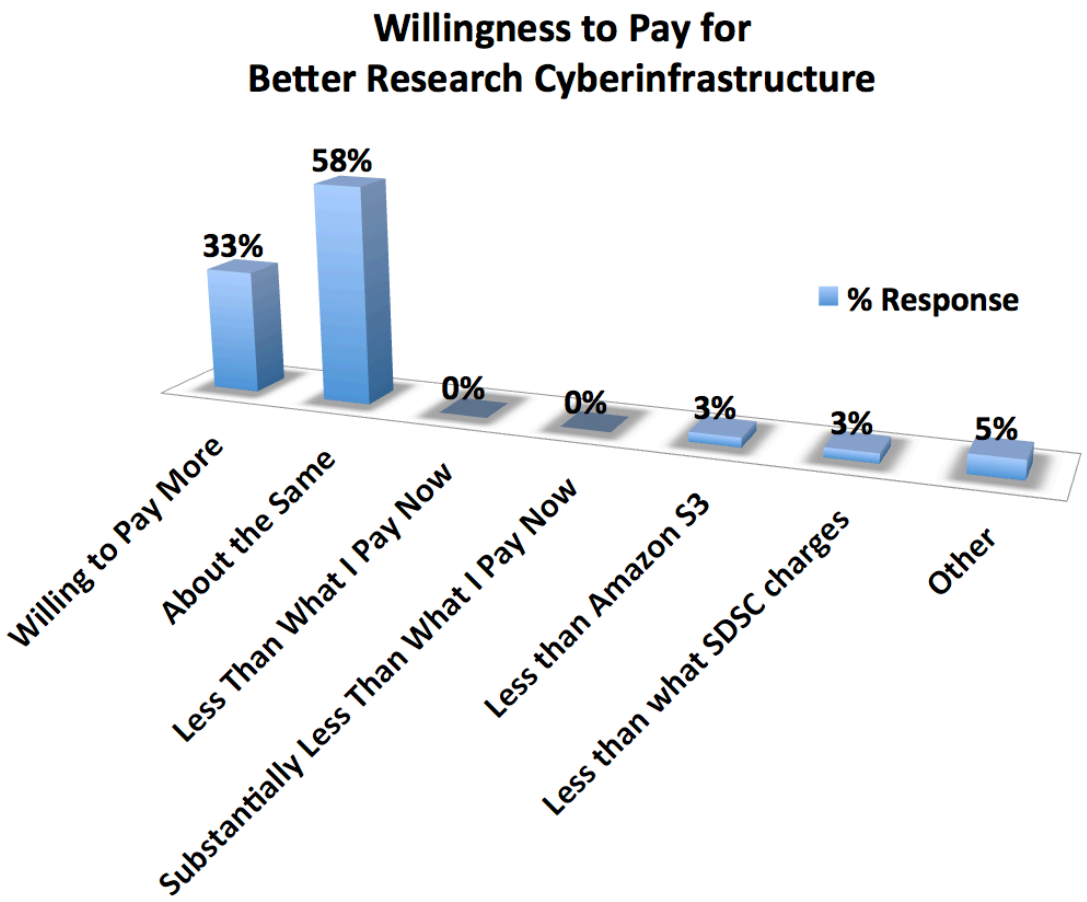
How long do you need to retain your data?

Interview Responses: Common Data Storage, Network and Security Requirements

Requirements	%
Data storage, management and sharing using NFS or CIFS	73%
Replicated backup/Disaster Recovery	66%
Need for 10 GbE or better Connection in lab	38%
SDSC as 2nd backup site, aka, 2nd copy site	18%
Dual site replication on campus	16%
Desktop workstation Backup	18%
Compliant/secure Storage	16%
Tiered storage media and prices	16%
Uniform campus wide UID/GID	10%

- **Reliable, NFS/CIFS storage most common**
- **Many responses relate to data durability – backups/copies/tiered storage**
- **High-speed networking**
- **“Compliant” environment (storage/computing)**

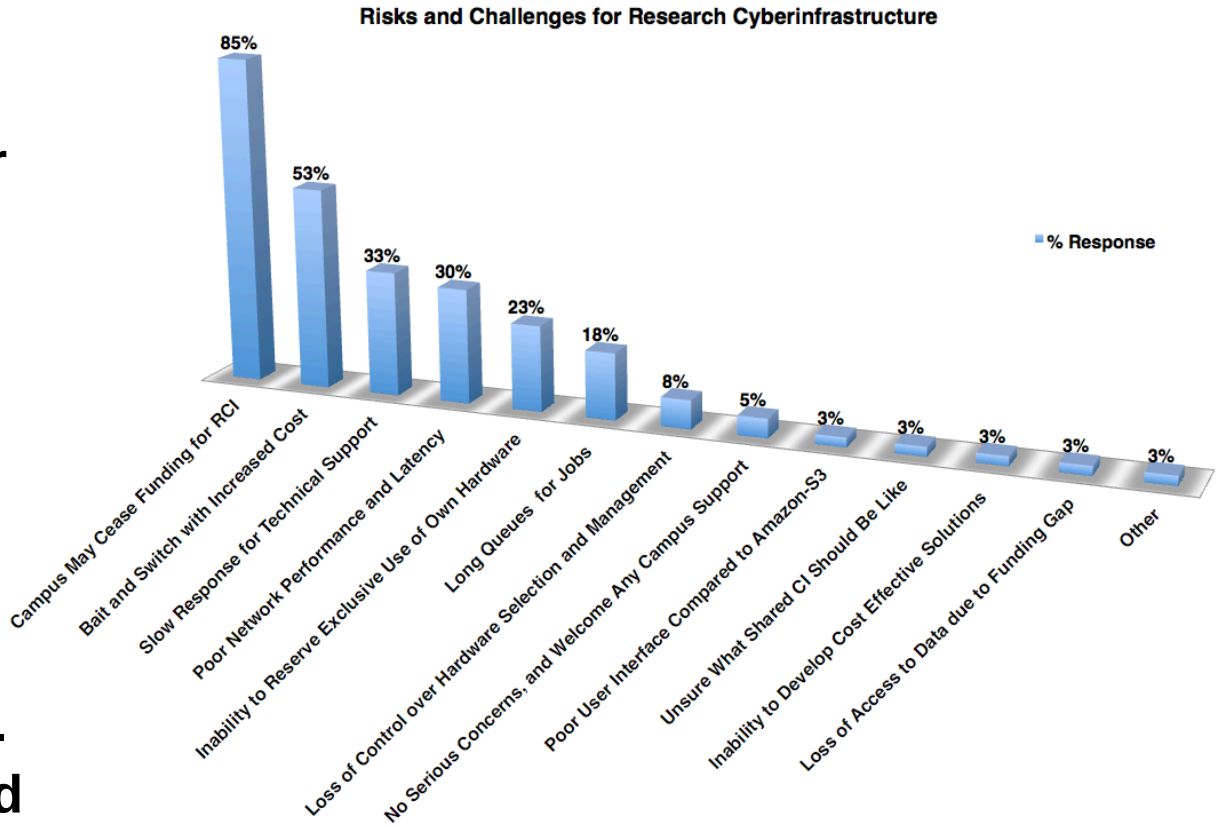
Interview responses: What are you willing to pay for storage?



- **Intentionally posed a relative question rather than absolute $'s**
- **Willing to pay "about the same" or "more" for shared services**
- **PIs comfortable with hardware costs but …**
- **Often ignore staff cost to operate equipment**
  - "Hard to compete with (perception of) slave labor" (Jim Pepin)

Interview responses: Risks and Challenges for Research Cyberinfrastructure

- **Skepticism re long-term campus commitment to RCI, or increased costs down the line, are major concerns**
  - Technical issues less important
- **Crucial that campus demonstrate commitment**
- **Potential chicken-and-egg between sustained commitment and adoption!**



Top 6 responses above correspond to options presented in survey (then "Other" write-in)

PI Interview Quotes: What Researchers Desire for Their RCI

- **Human expertise**
  - "Critical mass of technical knowledge on computing, networking, and storage to run the facility at the state of the art level. ***Consulting services, particularly critical for young investigators.*** Technical webinars, recorded and archived, would be very useful for distributed campus units"
  - "Help design the primary data stores, fast scratch space for intermediate analysis, hosted web space for final results, and backups of everything. Connectivity to local cluster, campus and national resources, and commercial cloud services."
- **Commercial Quality and Ease of Use**
  - "***Amazon S3 level stability; Campus wide Active Directory (AD) support; Dropbox-like user interface for users***"
  - "Tiered service/cost levels to provide the minimum, mid to full experience."
  - "Configure service offerings in ways that can be used as matching funds in grant applications."
- **Infrastructure services**
  - "Centralized services (virtualized compute/storage) at competitive price/performance."
  - "Great to have unlimited run times on my own cores, without having to resubmit."

PI Interview Quotes: What is RCI to Me?

- **Cost**
  - "***Dependable, economically reasonable storage for data protection***; Current SDSC Cloud/Project Storage are too expensive; Pricing models close to hardware cost plus part time labor needed"
  - "Campus may be required to provide matching funds for equipment donations and external funding. Shared internal resources are difficult to get commitments from donors."
  - "Love to find ways to archive data for long term that is cost effective and efficient, which do not need to be online most of the time."
- **Grants**
  - "***Cost-effectiveness is key in national competitions for grants***. Universities with large shared resources that provide additional opportunistic resources increase their competitiveness."
  - "***Help to keep costs within constraints of grant budget, and free research personnel/developers to do better things***."
- **Metadata**
  - "Where is the data from my student who just left?"
  - ***"Who else is working on this gene on campus?"***

Requirements-Based Storage: Initial Recommendation

- **Key Requirements from Interviews**
  - Build RCI services that the researchers need today
  - Build critical mass and expand on value-added services with feedback from community
- **Still sorting out all the requirements/recommendations from interviews, but one straightforward service emerged as a high priority for researchers**
  - Storage that is accessible, high performance and sustainable.
  - NFS, CIFS mounts from across campus, SDSC colo, TSCC
  - Replicated for durability and high availability
  - Simplified and flexible data storage provisioning at the school/department/lab levels
  - Professionally managed
  - Cost effective and sensitive to today's funding climate
- **Recommendation: *RCI Data Services (RDS)-Network Attached Storage***
  - Follow principles behind condo computing: PI buys hardware and pays an RCI-supplemented operations cost sufficient to cover incremental costs
  - Low operating cost, achieved through economy of scale, and wide adoption

Data Life Cycle Services for Campus Researchers

- **Initial Services (currently available)**
  - High-performance Lustre PFS for HPC scratch and medium-term parking
  - Centralized network-attached storage (NFS, CIFS mounts) available from campus HPC cluster and remotely across campus (10 Gbps infrastructure)
  - Support to Data Management Plans
- **Next-phase services**
  - Low-cost backup services for research data
  - Improve interfaces for cloud storage to facilitate sharing/access
  - HIPAA/PII protected data services
  - Data management tools and integrated solutions to DMPs
  - Build on experience with pilots to develop basic data curation services
- **Looking ahead**
  - Tools/technology to support data sharing (via SDSC Cloud, portals, gateways)
  - Develop tools for search, discovery and sharing of data
  - Integrate data management practices into routine research practices

What's Changed Along the Way for RCI?

- **Technical expertise remains central to RCI's objectives, but it's harder to 'package' into a scalable service**
- **Should view storage and curation in the context of an integrated user-centric data life cycle/research workflow**
- **Emergence of NSF's DMP and NIH's data sharing requirements, and recent OSTP guidance, impacts what researchers need now and in future**
- **Can campus support a dedicated research network in addition to shared campus network infrastructure?**
  - Issues of grant-funded networking capability (or other resources)
- **Colocation services should be viewed campus-wide, with consistent costs and incentives**
  - Very different facilities and costs; RCI offer a range of options?

RCI Lessons Learned

- **Tough budget environment for new initiatives – even ones with financial ROI and general technical support**
  - Substantial delays in getting budget approved and program moving forward
- **We should have asked for a multi-year budget commitment at outset**
  - PIs need to be confident program will be there; 'condo' programs impossible
- **Campus needs to subsidize total cost of ownership in order to overcome a natural preference for autonomy and control**
  - PI's often have different views of the total cost-of-ownership
- **Incentives need to be consistent**
  - e.g. Using closet down hall is free to PI, while energy-efficient colo costs $
- **Adoption takes time (marketing down, managing up)**
  - Demonstrated commitment and persistence and consistency (including costs)
  - Hearing about service from colleagues has more credibility than providers
  - Get cost of services into new grant proposals, not redirecting current budgets

How to get more info about RCI

- **Web site: rci.ucsd.edu**
- **For each production service, site includes**
  - Description of services
  - Cost summary
  - Approved text for PIs to use in proposals
- **Email rci@ucsd.edu**
- **Call Richard Moore, RCI Proj Mgr, ext 858-822-5457**